7.0 DATA ANALYSIS AND STATISTICAL TREATMENT

Good data analysis and statistical treatment practices are essential for the production of quality results from the effluent monitoring and environmental surveillance program required by DOE 5400.1 and DOE 5400.5. The goals for analyzing effluent monitoring and environmental surveillance data *should* be

- To estimate radionuclide concentrations at each sampling and/or measurement point for each sampling and/or measurement time, and estimate accuracy and precision
- To compare the estimated radionuclide concentrations at each sampling and/or measurement point to previous concentration estimates at that point to identify changes or inconsistencies in radionuclide levels
- To compare the radionuclide concentrations at each sampling and/or measurement point to the established limit(s), or concentrations related to the applicable dose limit, for those radionuclides
- To compare radionuclide concentrations at single sampling and/or measurement points or groups of points to those at control or other points and evaluate the reliability of those comparisons.

The statistical techniques used to support the concentration estimates, to determine their corresponding measures of reliability, and to compare radionuclide data between stations and times *should** be designed with consideration of the characteristics of effluent and environmental data. These characteristics include a time series of data with skewed distributions (usually lognormal), a high degree of variability, and often large amounts of missing data and readings that are below the detection limit of the sample analysis technique. Documented and approved sampling, sample-handling, analysis, and data-management techniques *should** be used to reduce variability of the results as much as possible. Data generated by the effluent monitoring and environmental surveillance program form the bases from which site management decisions are made. Thus, adequate attention to estimating the accuracy and precision of the data is necessary to determine whether such management decisions and actions are supported by valid and reliable data.

7.1 SUMMARY OF DATA ANALYSIS AND STATISTICAL TREATMENT REQUIREMENTS

The data analysis and statistical treatment procedures that are required to be incorporated into the radiological effluent monitoring and environmental surveillance program at a DOE site are presented in the summary. The level of confidence in the data due to the radiological analyses *should** be estimated by analyzing blanks and spiked pseudosamples and by comparing the resulting concentration estimates to the known concentrations in those samples. The precision of radionuclide analytical results *should** be reported as a range, a variance, a standard deviation, a standard error, and/or a confidence interval. Analytical precision estimates for radiological analyses *should* be made from replicate samples. Data *should** be examined and entered into the appropriate data bases promptly after analysis. When selecting the data to be considered, outliers *should** be excluded from the data only after investigation confirms that an error has been made in the sample collection, preparation, measurement, or data analysis process. As each data point is collected, it *should** be compared to previous data, because such comparison can help identify unusual measurements that require investigation or further statistical evaluation.

7.2 VARIABILITY OF EFFLUENT AND ENVIRONMENTAL DATA

The variability of the effluent data will determine the degree of precision and accuracy that can be achieved with the results. Careful design and execution of the monitoring program can substantially improve the quality of the effluent monitoring and environmental surveillance results.

7.2.1 Sources of Variability in Effluent/Environmental Data

The sources of variability in effluent data are listed in Table 7-1. These sources can be divided into three types: environmental, sampling, and recording. The analyses performed to determine and reduce the sources of variability *should* consider the relevancy of the variability source with respect to the actual conditions at the sampling and/or measurement point.

7.2.2 Estimating Accuracy and Precision

An estimate of the levels of accuracy and precision required for the data, based on previous site monitoring and surveillance experience, *should* be used to develop data analysis and handling strategies for the effluent monitoring and environmental surveillance programs. These strategies *should* then be re-evaluated periodically (or after significant modification to site conditions) to determine whether they are adequate for the present site conditions.

7.3 SUMMARIZATION OF DATA AND TESTING FOR OUTLIERS

Often, a measure of central tendency is needed to summarize the information in a data set (e.g., in the calculation of a yearly average concentration). In addition, an estimate of precision is required for that summary statistic. Assumptions about the underlying data distribution are inherent in the calculation of most statistical parameters; therefore, the distribution of the radionuclide concentration data *should* be established before the calculated parameters are considered valid.

7.3.1 Distribution Analysis

The assumption of a normal data distribution is implicit in the calculation of most statistical parameters. Radionuclide distributions are typically

TABLE 7-1. Sources of Variability in Effluent Monitoring Data (DOE/EP-0023)

Type	Source	Examples
Environmental	Space	Distance from emission source, elevation, heterogeneous dispersion of material
	Time	Variation in rates of emissions, variation in rates of dispersion
	Space x Time	Nonstationary differences between sampling stations over time
Sampling	Sample Collection	Nonrepresentative sampling, inconsistent sampling techniques, sampling equipment failure
	Sample Handling	Chemical reactions, nonuniform storage conditions, container effects
	Sample Processing	Volume or weight measurement errors, insufficient sample mixing, nonrepresentative subsampling
	Measurement	Calibration errors, instrument errors, readout errors
	Cross-Contamination	Residual contamination of containers and work areas, imperfect sealing of containers for transport, surface contamination from transport, separation of high- and low-activity samples, decontamination practices
Recording	Data Recording and Transfer	Errors in data entry, errors in transfer of data from lab books to computer files

lognormal, and when appropriate, the raw data *should* be transformed to logarithms before calculating summary statistics.

Data sets with more than 10 points *should* be tested for normality. (Data sets containing fewer than 10 points can be treated as either normal or lognormal.) The simplest and most straightforward test involves plotting the data points on commercially available normal or lognormal probability paper. If the data form an approximately continuous straight line, it can be concluded that the data are homogeneous and from a distribution of the same type as the probability paper (normal or lognormal) on which they are plotted.

Severe discontinuities in the straight line indicate that different subsets of the data, coming from different distributions, are involved. When such conditions occur, the data *should* be reexamined and identifiable subsets analyzed separately. Curvilinearity in the plot indicates that a data transformation is required before statistics based on the normal distribution are calculated.

Other acceptable methods of assessing normality are to plot all of the data in a frequency distribution and perform a χ^2 test for normality, or to visually inspect a histogram of the data. The method of assessing normality *should* be presented in reports of the data.

7.3.2 Measures of Central Tendency

The appropriate measure of central tendency depends on the characteristics of the radionuclide concentration data collected. For normally distributed data with only a small number of extreme or less-than-detectable values, the arithmetic mean is the appropriate estimator of central tendency. When the data set contains large numbers of extreme values or concentrations below the analytical detection limits, the median, which is less sensitive to extreme values than the mean, *should* be used to summarize the data. Trimmed means (arithmetic means calculated while excluding some percentage of the upper and lower data values) can also be appropriate in these cases.

The data *should* be transformed to approximate a normal distribution before the central values are calculated. Most often a log transformation will normalize environmental data.

The mean of a distribution can be read from a plot of the data on probability paper. The mean (which in the case of the normal distribution is equal to the median) is the 50th percentile intercept on the probability plot.

7.3.3 Measures of Dispersion

Dispersion in normally distributed data, without large numbers of outliers and less-than-detectable values, *should* be represented as a variance, a standard deviation, a standard error, or a confidence interval. Again, data *should* be transformed if necessary to approximate a normal distribution.

For data with substantial numbers of extreme values, other measures *should* be used to estimate the dispersion around the central value. The full range of data values or the interquartile range (the range of data between the 25th and 75th percentiles) and the median absolute deviation (the median of the differences between each data point and the indicator of central tendency) are also acceptable measures.

The slope of the line drawn through the data points plotted on probability paper is the standard deviation of the data.

7.3.4 Less-Than-Detectable Values

Monitoring programs often include measurement of extremely low concentrations of radionuclides, below the detection limit of the counting instruments. Data sets with large numbers of less-than-detectable values need special consideration in the statistical analyses (Gilbert 1987).

Less-than-detectable data will produce numerical measurements with values below the detection limit and sometimes negative values. All of the actual values, including those that are negative, *should* be included in the statistical analyses. Practices such as assigning a zero, the detection limit value, or some in-between value to the below-detectable data point, or discarding those data points can severely bias the resulting parameter estimates and *should* be avoided.

When analytical instruments or laboratories do not supply the actual values for readings less than the detection limit, but make some designation such as "ND," the actual values for those data points *should* be obtained. When obtaining these data points is not possible, at least the number of less-than-detectable values *should* be obtained. Data from censored distributions (for which the number of less-than-detectable values is known) are more amenable to standard statistical analyses than are those from truncated distributions (for which the number of values below the detection limit are not known), which require special statistical techniques (Gilbert and Kinnison 1981).

7.3.5 Testing for Outliers

An outlier is defined as an abnormally high or low data value. It can represent a true extreme value, or it can indicate data errors or equipment malfunctions or errors. It is important to compare each data point to previous data to determine whether the point is an outlier or a true data point that is to be included in the data set (Gilbert 1987).

Several statistical tests are available to test for outliers. Most of these tests assume a normal distribution, so data *should* be transformed to approximate the normal distribution before outlier tests are performed. Outliers can be identified qualitatively by adding the new data point to the data probability plot and noting if the point falls on an extreme end of the plot line; alternatively, a 2- or 3-standard-deviation probability ellipse can be constructed around a scatterplot of all of the data, with points falling outside of that ellipse considered outliers. These tests, while statistically valid (as long as their assumptions, e.g., normality, are met), determine only whether the new point is extreme with respect to the mean or median of the entire data set and do not detect temporal irregularities (for example, data values that are close to a yearly average but highly unusual for the season or time of day at which they occurred). Therefore, these tests are not adequate to serve as the sole justification for the inclusion or exclusion of data from the set. A better procedure that takes into consideration the temporal pattern of the observations involves the development of a time plot of the data, with each new data value being entered promptly after collection. Outliers can be identified by inspection of the time plot. Control charting can also be a useful technique for identifying outliers. Control charts are time plots on which the center line represents the mean or median concentration value, and 1-, 2-, and 3-standard-deviation bands are marked. Data points falling outside of the 2- or 3-standard-deviation confidence bands are considered outliers. The position of the center line can differ diurnally, seasonally, or yearly. The central values *should* be calculated separately for identified subgroups of the data. Control charting is not useful for some new monitoring programs because they require sufficient amounts of data to adequately estimate the mean value and standard deviation for each subgroup. Graphs of moving averages of the data *should* also be plotted for each station, as soon as sufficient amounts of data (at least 10 points) are acquired. These plots will indicate overall trends in the data, identification of which aids in data interpretation as well as in detecting sampling or equipment errors.

When outliers are identified, a decision must be made whether to include those numbers in estimates of radionuclide concentrations or in comparisons between data sets. Outliers can represent true extreme values or can indicate malfunctions or failures in sampling equipment or variability in sample quality. Most often what at first appear to be outliers prove to be data transcription errors. The presence of outliers can, however, severely affect the value of the estimated mean or the outcome of statistical comparisons. When outliers that are not attributable to errors are contained in the data set, estimators and statistical tests *should* be computed with and without the outliers to see if the results of the two calculations are markedly different. If the results differ substantially because of outliers in the data, then both results *should* be reported.

7.3.6 Elements of Good Practice

Certain procedures *should* be followed that will aid in the interpretation of the effluent monitoring data and improve the quality of the results from the program by helping to detect erroneous measurements. Comments on the quality of the samples taken *should* be entered into the data base with the sample radionuclide concentration measurements. In addition to the data collected during the regular sampling program, logs of events that might affect radionuclide concentrations (e.g., precipitation) *should* be kept.

7.4 TREATMENT OF SIGNIFICANT FIGURES

Often, calculations involving measured values result in numbers with more significant figures than were in the original measurements and give an erroneous impression of the precision and accuracy of results. The number of significant figures in reported data *should* reflect the precision of the measured values. A larger number of figures may be carried during the calculations for computational accuracy. The number of significant figures reported for raw data *should* reflect the true precision of the measurement technique.

When measurements are multiplied or divided, the number of significant figures in the product or quotient *should* not exceed that of the least precise measurement used in the calculations. When measurements are added or subtracted, the recorded precision of the result *should* not exceed that of the least precise measurement.

7.5 PARENT-DECAY PRODUCT RELATIONSHIPS

A common practice in the monitoring of radionuclide concentrations is to measure the activity of the parent radionuclide and calculate the amount of the decay products present from the known physical relationships. As an alternative, the concentrations of parent nuclides may be calculated from the measurement of the decay products. These calculations are relatively straightforward when the parent and decay products are at equilibrium, and in the absence of contrary data. Corrections *should* be made for calculations performed during the transitory period before equilibrium is reached. Correct estimation of the amount of the decay product (or parent) material present requires definite knowledge of the difference between the time of measurement and the time of the initiation of parent decay. The recorded accuracy and precision of the calculated radionuclide concentration estimates, as indicated by number of significant figures, *should* not exceed those of the original measured concentration. Uncertainties in the length of time between measurement and the initiation of parent decay *should* be reported and incorporated into the precision estimates for the calculated concentrations.

7.6 COMPARISONS TO REGULATORY OR ADMINISTRATIVE CONTROL STANDARDS AND CONTROL DATA

The object of obtaining reliable estimates of radionuclide concentrations at the monitoring stations is to compare those values to regulatory or administrative control standards or values at control stations to determine whether action must be taken to reduce the radionuclide levels in the effluents.

7.6.1 Single Concentration Measurements

Statistical tests are not appropriate for comparisons of single values, such as when a single radionuclide concentration measurement is compared to its regulatory limit. Single values can have a large associated uncertainty, and they are not necessarily an accurate representation of how well the facility is complying with the limit. Thus, additional sampling and/or measurement *should* be considered to provide an accurate representation of compliance status.

7.6.2 Groups of Measurements

Concentration estimates from groups of sampling and/or measurement points *should* be compared using standard (parametric) analysis of variance techniques (Winer 1971) when the data meet the underlying assumptions of those tests. Standard nonparametric statistical comparison techniques (Hollander and Wolfe 1973) *should* be used when the assumptions of the parametric tests are not met by the data. Caution *should* be used when comparing groups of readings from single points over time, because of the likely strong autocorrelation in the time series of data.

7.7 QUALITY ASSURANCE

As they apply to data analysis and statistical treatment activities, the general quality assurance program provisions of Chapter 10 *should** be followed. Specific quality assurance activity requirements for data analysis and statistical treatment activities at a site *should* be incorporated in the Quality Assurance Plan for the facility.